International Scientific Technical and Economic Research (ISSN: 2959-1309)
Vol. 4, No. 1, (March 2026), pp. 200-221
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE DOI: 10.71451/ISTAER2610

A Deep Reinforcement Learning Signal Control Algorithm for Traffic Carbon Emission Optimization

Hanyu Xu*

Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China

Abstract: Urban traffic congestion leads to frequent vehicle start-stop events and low-speed operation, which is one of the primary drivers of carbon emission growth. To address the problems of multi-objective conflict, training instability, and inadequate carbon emission modeling in existing traffic signal control methods for carbon emission optimization, this paper proposes a deep reinforcement learning signal control algorithm for carbon emission optimization. This method constructs a carbon-emission-aware dynamic reward mechanism and achieves collaborative optimization of traffic efficiency and emission reduction objectives through adaptive weight adjustment; Lagrange multiplier method is introduced to embed the carbon emission threshold as an explicit constraint into the strategy learning process to ensure that the emission level is controlled within an acceptable range; For multi-intersection scenarios, a distributed collaborative control framework based on parameter sharing and neighborhood information interaction is designed to enhance the model's ability to perceive the spatial propagation characteristics of traffic flow. Based on the SUMO simulation platform, experimental validation is conducted in three scenarios: a single intersection, a 4×4 grid network, and a real-world urban road network. The results show that compared with PPO algorithm, the average carbon emissions of this method are reduced by 11.3% to 12.8%, average delay is reduced by 15.7%, average speed is increased by 9.6%, and the comprehensive performance index is improved by 12.2%; During the training process, the fluctuation of strategy is reduced by about 50%, and the degradation rate of generalization performance is reduced by 34.2% compared with the comparison method. This study provides an effective intelligent solution for low-carbon-oriented urban traffic signal control.

Keywords: Deep reinforcement learning; Traffic signal control; Carbon emission optimization; Multi objective optimization; Constraint reinforcement learning

How to Cite: Xu, H. (2026). A Deep Reinforcement Learning Signal Control Algorithm for Traffic Carbon Emission Optimization. *International Scientific Technical and Economic Research, 4*(1), 200–221. https://doi.org/10.71451/ISTAER2610

Article history: Received: 10 Jan 2026; **Revised:** 25 Feb 2026; **Accepted:** 23 Mar 2026; **Published:** 29 Mar 2026

1. INTRODUCTION

As urbanization accelerates, the number of motor vehicles continues to increase, and traffic

*Corresponding author: Hanyu Xu, Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China. Email: hanyuxu3-c@my.cityu.edu.hk

congestion has become one of the core issues restricting sustainable urban development. Traffic congestion not only leads to a decline in traffic efficiency but also causes a significant increase in carbon emissions due to frequent vehicle start-stop events and low-speed operation [1],[2]. Research shows that carbon emissions from the urban transport sector account for nearly a quarter of global total emissions, and their growth rate is far higher than that of other sectors. This phenomenon reveals the profound coupling between traffic congestion and carbon emissions: congestion exacerbates emissions, while the accumulation of emissions places pressure on the ecological environment, forming a mutually reinforcing negative feedback loop [3]. Therefore, how to effectively reduce carbon emissions while improving traffic operational efficiency has become a key scientific problem in the field of traffic management and control.

Traditional traffic signal control methods, such as fixed-time control and actuated control, are mostly based on historical statistics or simple rule-based logic, which struggle to cope with the high dynamics and uncertainty of traffic flow. In recent years, deep reinforcement learning has shown significant potential in traffic signal control due to its strong perception and decision-making capabilities. Through continuous interaction between the agent and the environment, such methods can adaptively learn optimal control strategies and effectively improve intersection traffic efficiency [4],[5],[6]. However, most existing studies take minimizing delay or queue length as the optimization objective, ignoring the direct impact of signal control strategies on vehicle operating status and carbon emissions. Although some efforts have begun to take emissions into account, they mostly remain at the level of post-hoc evaluation, lacking an effective mechanism to embed the carbon emission indicator system into the reinforcement learning training process.

At present, deep reinforcement learning faces multiple challenges in traffic carbon emission optimization. First, the instantaneous and cumulative requirements of carbon emission models must accurately describe the relationship between vehicle dynamics and emissions, which imposes higher requirements on state representation and reward design. Second, the reinforcement learning training process itself suffers from slow convergence and large fluctuations. In multi-objective conflict scenarios, how to balance the trade-off between emissions and efficiency and improve training stability becomes the key to the practical application of the algorithm. In addition, most existing methods use fixed weights in handling carbon emission constraints, which makes it difficult to adapt to changes in traffic conditions and can easily lead to policy deviation or performance degradation.

To address the above problems, this paper proposes a deep reinforcement learning signal control algorithm for traffic carbon emission optimization, aiming to achieve collaborative optimization of emissions and efficiency. The main contributions of this paper include: first, a multi-objective carbon-emission-aware reward mechanism is constructed, which is combined with a dynamic weight adjustment strategy, enabling the model to adaptively shift its optimization focus according to traffic conditions and effectively alleviate multi-objective conflicts; Second, a deep reinforcement learning framework integrating constrained optimization is designed, and the carbon emission threshold is explicitly incorporated into the policy learning process using the Lagrange multiplier method to ensure that emission levels are always kept within an acceptable range; Third, a distributed collaborative control mechanism for multi-intersection scenarios is proposed. Through parameter sharing and neighborhood information interaction, the model's ability to perceive the spatial propagation characteristics of traffic flow is enhanced, and the collaborative control effectiveness of the overall road network is improved. Through the above innovations, this paper provides a systematic solution for realizing low-carbon-oriented intelligent traffic control.

2. RELATED WORK

Traffic signal control methods have evolved from traditional control to intelligent control. Early fixed-time control presets signal timing schemes based on historical traffic flow data.

Although it is simple to implement, it cannot cope with the dynamic changes in traffic flow. Actuated control obtains lane occupancy in real time via deployed detectors and triggers phase switching according to a preset threshold, which improves adaptability to a certain extent, but its rules remain essentially reactive, lacking the ability to predict overall traffic conditions. With the improvement of computing power, methods based on model predictive control have been applied to traffic signal optimization [7],[8]. Signal timing is dynamically adjusted through rolling optimization, demonstrating good control performance. However, the performance of these methods is highly dependent on the accuracy of traffic prediction models, and they face the problem of high computational complexity in large-scale road networks [9],[10]. In general, traditional control methods have obvious limitations in dealing with high-dimensional dynamic traffic environments, and it is difficult to achieve refined multi-objective collaborative optimization.

The rise of deep reinforcement learning has opened a new research paradigm for traffic signal control. In this approach, the signal control problem is modeled as a Markov decision process, and the optimal control strategy is automatically learned through interaction between the agent and the simulation environment [11],[12]. In single-intersection scenarios, deep Q-networks and their variants are widely used for phase selection, demonstrating the effectiveness of deep reinforcement learning in reducing vehicle delay and queue length [13]. As research deepens, policy gradient methods such as proximal policy optimization have gradually become the mainstream choice due to their superior continuous control capability and training stability. In multi-intersection collaborative control, researchers attempt to use the multi-agent reinforcement learning framework to alleviate environmental non-stationarity through centralized training with decentralized execution. In some works, graph neural networks are introduced to model the topology of the road network, which improves the agent's ability to perceive neighboring traffic states [14]. However, most existing studies take traffic efficiency as the core optimization objective and treat emissions as a secondary indicator for post-hoc analysis, lacking deep integration with the mechanisms of carbon emission generation.

Research on traffic carbon emission estimation models provides fundamental support for incorporating emission targets into control optimization. Macro-level emission models estimate total emissions based on average speed or mileage, which are simple to compute but fail to reflect the impact of microscopic driving behavior on emissions [15]. Micro-level emission models can effectively capture the contribution of acceleration and deceleration processes to emissions by finely modeling instantaneous emissions based on vehicle speed, acceleration, and other operating parameters. At the meso level, methods based on vehicle-specific power correlate engine load with emissions, balancing accuracy and computational efficiency to a certain extent [16]. At present, most of these emission models are used to assess the environmental impact of different traffic management measures, while attempts to directly serve as optimization objectives in signal control strategy generation remain relatively limited. How to organically integrate the emission model with the reinforcement learning training framework so that it can guide policy updates in real time is a key challenge that current research must address.

In the field of multi-objective and constrained reinforcement learning, existing research provides methodological references for handling the trade-off between carbon emissions and traffic efficiency. Multi-objective reinforcement learning seeks to find a balance between multiple objectives by constructing Pareto frontiers or designing compound reward functions [17],[18],[19]. The traditional approach is to convert the multi-objective problem into a single-objective problem via linear weighting, but the weight setting often depends on experience, making it difficult to maintain optimality when traffic conditions change [20],[21]. For constrained optimization problems, the Lagrange multiplier method is widely used in constrained reinforcement learning. By transforming constraints into penalty terms and dynamically adjusting the penalty coefficient, the policy can maximize cumulative rewards while satisfying the constraints. In recent years, scholars have begun to focus on the field of safe reinforcement learning, treating environmental constraints such as emissions as safety

boundaries and ensuring that the system operates within an acceptable range through the constrained policy space [22],[23]. These methods have laid a theoretical foundation for the development of low-carbon-oriented signal control algorithms, but how to effectively combine them with the spatiotemporal dynamic characteristics of carbon emissions in traffic scenarios still requires further exploration.

Based on the above research progress, there remain obvious deficiencies in the application of deep reinforcement learning to traffic signal control. First, existing methods generally regard carbon emissions as a secondary indicator and lack a mechanism to embed the emission constraint system into the policy learning process, making it difficult to achieve low-carbon objectives in the optimization results. Second, in dealing with multi-objective conflicts, the traditional fixed-weight method struggles to adapt to dynamic changes in traffic conditions, which can easily cause policy deviation or performance degradation. Third, in multi-intersection collaborative control, existing models do not adequately describe the spatial propagation characteristics of traffic flow, which limits the emission reduction effectiveness of the overall road network. Fourth, there are problems in the algorithm training process, such as large deviations in value estimation and drastic fluctuations in policy, which affect the reliability of actual deployment. To address the above shortcomings, this paper proposes a deep reinforcement learning signal control algorithm for carbon emission optimization, which aims to achieve collaborative optimization of traffic efficiency and carbon emissions by constructing a carbon-emission-aware dynamic reward mechanism, introducing a constrained optimization framework, and designing a multi-intersection collaborative control mechanism.

3. METHODOLOGY

3.1 System Architecture

This paper constructs a deep reinforcement learning-based signal control framework for traffic carbon emission optimization. Its overall architecture consists of three parts: a traffic simulation environment, a reinforcement learning agent, and a carbon emission estimation module. The traffic environment depicts vehicle behavior through a microscopic simulation platform and outputs real-time status, including vehicle flow dynamics, speed changes, and queue information; The reinforcement learning agent generates signal control strategies based on the current state; The emission estimation module calculates instantaneous carbon emissions based on the vehicle operating state and provides feedback to the learning process to guide policy updates.

At time step t, the system state is denoted as s_t, and the agent generates action a_t according to the policy $\pi_\theta(a_t \mid s_t)$, where θ represents the policy network parameters. After the action is applied to the traffic environment, the environmental state transitions to s_{t+1}, and the emission estimation module calculates the carbon emission E_t. The system reward r_t is composed of emissions and traffic efficiency, and is used to update the policy parameters. This process forms a standard Markov decision process (MDP) closed loop [24]:

$$s_t \xrightarrow{a_t} s_{t+1}, r_t = f(E_t, D_t) \tag{1}$$

Where D_t is the traffic delay metric. Through continuous interaction, the system gradually learns the optimal signal control strategy to realize the collaborative optimization of carbon emissions and traffic efficiency.

3.2 State & Action Representation

To fully describe the complex traffic environment, this paper designs a multi-dimensional traffic state representation vector. For intersection i, its state at time step t is defined as:

$$s_t^i = \{q_t^i, v_t^i, \rho_t^i, w_t^i\} \tag{2}$$

Where q_t^i is the queue length per lane, v_t^i is the average speed vehicle, ρ_t^i is the traffic density, and w_t^i is the cumulative waiting time. These characteristics can describe the traffic operation state from different angles.

Considering the spatial correlation between intersections, neighborhood state information is introduced to construct a multi-scale state representation [25]. For intersection i, its extended status is:

$$\tilde{s}_t^i = s_t^i \oplus \sum_{j \in \mathcal{N}(i)} \alpha_{ij} s_t^j \tag{3}$$

Where $\mathcal{N}(i)$ is the set of neighboring intersections, α_{ij} is the adjacency weight, and $\oplus$ is the feature concatenation. The design enhances the model's perception of traffic propagation effects.

The action space is defined as the decision variables of the signal control strategy. This paper adopts a joint modeling approach combining discrete phase control with continuous timing:

$$a_t = (p_t, \tau_t) \tag{4}$$

Where p_t represents the selected signal phase and τ_t represents the phase duration. This design considers both control flexibility and implementation feasibility.

3.3 Carbon aware Reward Design

To effectively embed the carbon emission objective into the reinforcement learning process, this paper constructs an emission estimation model based on vehicle operating states. The carbon emission of an individual vehicle at time step t can be expressed as [26]:

$$e_t = \alpha_1 v_t + \alpha_2 a_t + \alpha_3 v_t^2 \tag{5}$$

Where v_t is the vehicle speed, a_t is the acceleration, and $\alpha_1, \alpha_2, \alpha_3$ are model parameters. The total emission at the intersection is the sum of all vehicle emissions:

$$E_t = \sum_{k=1}^{N_t} e_t^k \tag{6}$$

Where N_t is the number of vehicles at the current time step.

On this basis, a multi-objective reward function is constructed:

$$r_t = -\lambda_E E_t - \lambda_D D_t \tag{7}$$

Where λ_E and λ_D are the weight coefficients for carbon emissions and delay respectively, and D_t is the average delay. To alleviate multi-objective conflicts, a dynamic weight adjustment mechanism is introduced:

$$\lambda_E^{t+1} = \lambda_E^t + \eta\left(E_t - E_{target}\right) \tag{8}$$

Where η is the adjustment step size and E_{target} is the target emission level. This mechanism enables the model to adaptively adjust the optimization focus under different traffic conditions, so as to improve the learning stability and strategy effectiveness.

3.4 Deep reinforcement learning algorithm design

This paper uses a deep reinforcement learning method based on the actor-critic framework to optimize the policy. The output action distribution $\pi_\theta(a \mid s)$ of the policy network (actor) and the estimated state value function $V_\phi(s)$ of the value network (critic), where θ and ϕ are the network parameters respectively.

To reduce value estimation bias, a dual value network structure is introduced:

$$V(s) = \min\{V_{\phi_1}(s), V_{\phi_2}(s)\} \tag{9}$$

This mechanism can effectively alleviate the over estimation problem and improve the stability of training.

For experience utilization, a prioritized experience replay mechanism is introduced. Sample priority is defined by the TD error [27]:

$$\delta_i = | r_i + \gamma V(s_{i+1}) - V(s_i) | \tag{10}$$

Sampling probability is:

$$P(i) = \frac{\delta_i^\alpha}{\sum_j \delta_j^\alpha} \tag{11}$$

Where, α controls the priority. This method improves the utilization efficiency of key samples.

In addition, an adaptive exploration strategy is introduced, and the entropy regularization term of the policy is dynamically adjusted:

$$L_{entropy} = -\beta_t \mathbb{E}[\log \pi_\theta(a \mid s)] \tag{12}$$

Where, β_t gradually decreases during training, to achieve the balance between exploration and exploitation.

3.5 Constrained Optimization

To further control carbon emission levels, the problem is formulated as a constrained reinforcement learning problem, with the objective:

$$\max_\pi \ \mathbb{E}\left[\sum_t r_t\right], \text{s.t.} \mathbb{E}[E_t] \le C \tag{13}$$

Where C is the emission constraint threshold.

The Lagrange multiplier method is used to transform the constrained problem into an unconstrained optimization [28],[29]:

$$\mathcal{L}(\theta, \lambda) = \mathbb{E}\left[\sum_t r_t\right] - \lambda(\mathbb{E}[E_t] - C) \tag{14}$$

Where λ is the Lagrange multiplier, which is updated as follows:

$$\lambda^{t+1} = [\lambda^t + \eta(E_t - C)]^+ \tag{15}$$

This method ensures that emission constraints are satisfied while optimizing the objective.

To enhance training stability, a constraint regularization term and a gradient clipping mechanism are introduced to smooth policy updates and avoid violent fluctuations.

3.6 Multi-intersection Coordination

For multi-intersection scenarios, this paper uses a distributed reinforcement learning framework in which each intersection acts as an independent agent making decisions, and collaborative optimization is achieved through parameter sharing. All agents share the policy network parameters θ, but make decisions based on their local states:

$$a_t^i \sim \pi_\theta(a \mid s_t^i) \tag{16}$$

To enable information interaction between intersections, a communication mechanism is introduced to encode the neighborhood state as a message vector:

$$m_t^i = \sum_{j \in \mathcal{N}(i)} W s_t^j \tag{17}$$

Where W is a learnable weight matrix. The final decision is based on the fusion state:

$$\hat{s}_t^i = s_t^i \oplus m_t^i \tag{18}$$

This mechanism can effectively capture the spatial propagation characteristics of traffic flow, thereby improving the coordination and control capability of the overall road network and enhancing carbon emission optimization.

4. ALGORITHM IMPLEMENTATION

To implement the above methodological framework, this paper constructs a deep reinforcement learning algorithm based on the actor-critic structure and carbon emission constraint optimization. The overall training process is carried out through continuous interaction with the traffic simulation environment. In each training episode, the agent generates signal control actions according to the current policy and updates the policy parameters using environmental feedback. Let the policy network be $\pi_\theta(a \mid s)$ and the value network be $V_\phi(s)$, where θ and ϕ denote the corresponding network parameters, respectively, and the discount factor be $\gamma \in (0,1)$, then the state value function is defined as:

$$V^\pi(s_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \right] \tag{19}$$

Where r_t represents the instantaneous reward at time step t. Based on this, the advantage function can be expressed as:

$$A_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{20}$$

The advantage function is used to guide the policy gradient update, thereby improving the efficiency of policy learning.

The core process of the algorithm can be formally described as follows: after initializing the policy network parameters θ, the value network parameter ϕ and the Lagrange multiplier λ, the process of environment interaction and parameter update is repeated. At each time step, the agent samples an action $a_t \sim \pi_\theta(\cdot \mid s_t)$,and receives the next state s_{t+1}, reward r_t and carbon emission E_t from the environment according to the current policy. The interaction data is stored in an experience buffer, and the parameters are updated in batches after each episode. The policy parameters are optimized by maximizing the objective function:

$$J(\theta) = \mathbb{E}[\log \pi_\theta(a_t \mid s_t) A_t] \tag{21}$$

The value network is updated by minimizing the mean square error:

$$L(\phi) = \mathbb{E}\left[\left(V_\phi(s_t) - y_t\right)^2\right] \tag{22}$$

Where $y_t = r_t + \gamma V_\phi(s_{t+1})$ is the target value. Combined with carbon emission constraints, the overall optimization objective is further revised to [30]:

$$J'(\theta) = \mathbb{E}[\log \pi_\theta(a_t \mid s_t) A_t] - \lambda(E_t - C) \tag{23}$$

Where C is the carbon emission threshold and λ is the dynamically updated Lagrange multiplier.

During training, the algorithm follows an iterative "interaction-storage-update" mode. Specifically, in each simulation episode, the system continuously collects state transition sequences $(s_t, a_t, r_t, s_{t+1}, E_t)$,and periodically updates the policy and value networks. To improve training stability, batch updating and a target network delay mechanism are introduced to make parameter updates smoother. In addition, by normalizing the advantage function:

$$\hat{A}_t = \frac{A_t - \mu_A}{\sigma_A} \tag{24}$$

Where μ_A and σ_A represent the mean and standard deviation of the advantage function respectively, to reduce the gradient oscillation and improve the convergence speed.

In terms of convergence and stability, this algorithm relies on the theoretical basis of policy gradient methods. If the learning rate α is sufficiently small and satisfies $\sum_t \alpha_t = \infty$, and $\sum_t \alpha_t^2 < \infty$, the policy parameter update can ensure convergence to a local optimal solution. In addition, by introducing the dual value network and constraint optimization mechanism, the issues of value estimation bias and policy oscillation are effectively mitigated. In the policy update process, the clipped objective function is used:

$$L^{clip}(\theta) = \mathbb{E}[min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\, A_t)] \tag{25}$$

Where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$, and ϵ is the clipping parameter. This mechanism limits the range of policy updates, thus significantly improving training stability.

In terms of complexity, assume that the parameter sizes of the policy network and the value network are $|\theta|$ and $|\phi|$ respectively, and the computational complexity of each forward and backward pass is approximately $O(|\theta| + |\phi|)$. If the number of sampling steps per episode are T and the batch size is B, the complexity of training per episode is:

$$O(T(|\theta| + |\phi|) + B(|\theta| + |\phi|)) \tag{26}$$

In a multi-intersection scenario, if there are N intersection agents and the parameter sharing mechanism is adopted, the overall complexity is approximately:

$$O\big(N \cdot T \cdot (\ |\theta| + |\phi|\ )\big) \tag{27}$$

The complexity increases linearly with the number of intersections, demonstrating good scalability. Through parallel simulation and distributed training, the actual computational overhead can be further reduced, making the algorithm feasible in large-scale traffic networks.

5. EXPERIMENTAL SETUP

To comprehensively verify the effectiveness of the proposed algorithm in traffic carbon emission optimization, this paper constructs a multi-scenario experimental environment based on a microscopic traffic simulation platform and integrates the emission model to evaluate policy performance. The experiment uses SUMO (Simulation of Urban Mobility) as the underlying simulation engine and interacts with the reinforcement learning framework in real

time through the TraCI interface. To ensure the reliability of the experimental results, an emission calculation model based on speed and acceleration is introduced to accurately estimate carbon emissions at each time step. Specifically, the instantaneous emission of vehicle k at time step t is defined as:

$$e_t^k = \beta_0 + \beta_1 v_t^k + \beta_2 (v_t^k)^2 + \beta_3 a_t^k \tag{28}$$

Where v_t^k is the vehicle speed, a_t^k is acceleration, $\beta_0, \beta_1, \beta_2, \beta_3$ are the empirical parameters. The total emission of the system is:

$$E = \sum_{t=1}^{T} \sum_{k=1}^{N_t} e_t^k \tag{29}$$

Where T represents the total number of simulation time steps, and N_t represents the number of vehicles at time step t. The model can describe the dynamic impact of traffic control strategies on emissions in a fine-grained manner.

For road network and traffic scenario settings, this paper constructs three typical traffic environments, including a single intersection, a 4×4 grid network, and a real-world urban road network (based on open-source map data) [31],[32]. The traffic flow is generated according to a Poisson distribution, and with arrival rate is λ. Different traffic intensities are achieved by adjusting λ. The vehicle generation process can be expressed as:

$$P(n) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{30}$$

Where $P(n)$ is the probability that the number of vehicles arriving per unit time is n. The experiment sets three traffic load levels: low, medium, and high, corresponding to $\lambda = 300,600,900$ veh/h respectively. Table 1 shows the specific configuration parameters for different experimental scenarios.

Table 1. Configuration of traffic scenarios and simulation settings

Scene type	Number of intersections	Number of lanes	Traffic intensity (veh/h)	Simulation duration (s)	Total number of vehicles
Single intersection	1	8	300	3600	1050
Grid road network	16	64	600	3600	4200
Urban road network	25	112	900	3600	8600
High load grid	16	64	900	3600	6300
Low load city	25	112	300	3600	2900

It can be seen from Table 1 that as the number of intersections expands from 1 to 25, the number of lanes and the total number of vehicles increase exponentially. For example, the number of vehicles in the urban road network scenario reaches 8600, which is more than 7 times higher than that in the single-intersection scenario. This scale expansion significantly increases the state space dimension and decision complexity. At the same time, under the same road network structure, increasing the traffic intensity (for example, from 600 to 900 veh/h) increases the total number of vehicles by about 50%, further aggravating congestion. Therefore, the experimental setup can effectively verify the robustness and scalability of the algorithm under high-dimensional state and high-load conditions.

For comparison, this paper selects a variety of representative baseline algorithms, including fixed-time control, actuated control, standard DQN, and PPO. All reinforcement learning methods are trained under the same environment and parameter settings to ensure fairness.

For evaluation metrics, the focus is on carbon emission performance while also considering traffic efficiency. Average carbon emissions are defined as:

$$\bar{E} = \frac{1}{T}\sum_{t=1}^{T} E_t \tag{31}$$

Average delay is defined as:

$$\bar{D} = \frac{1}{N}\sum_{k=1}^{N} (t_k^{actual} - t_k^{free}) \tag{32}$$

Where t_k^{actual} is the actual travel time of the vehicle, and t_k^{free} is the travel time under free-flow conditions. The average speed is defined as:

$$\bar{V} = \frac{1}{N}\sum_{k=1}^{N} v_k \tag{33}$$

For parameter configuration, the policy network and value network both adopt a three-layer fully connected structure, with 128, 128, and 64 neurons per layer, respectively. The learning rate is set to $\alpha = 3 \times 10^{-4}$, the discount factor $\gamma = 0.99$, and the batch size is 64. The initial value of the Lagrange multiplier is set to 0.1 and updated as follows:

$$\lambda_{t+1} = \max\big(0, \lambda_t + \eta(E_t - C)\big) \tag{34}$$

Where $\eta = 0.01$, and C is the emission threshold. With the above settings, the algorithm can quickly converge to an effective policy while ensuring training stability.

To sum up, the experimental design has been systematically designed from simulation environment, scene construction, comparison method and evaluation index, and the effectiveness and superiority of the proposed method in the optimization of traffic carbon emissions have been verified by multi-dimensional data.

6. RESULTS & ANALYSIS

To systematically evaluate the effectiveness of the proposed method, this section analyzes the overall performance, carbon emission optimization efficacy, training characteristics, and model generalization ability. First, we define the comprehensive performance index:

$$J = \omega_1 \bar{E} + \omega_2 \bar{D} - \omega_3 \bar{V} \tag{35}$$

Where $\bar{E}$ is the average carbon emission, $\bar{D}$ is the average delay, $\bar{V}$ is the average speed, and $\omega_1, \omega_2, \omega_3$ are the weight coefficients. This metric is used to evaluate the comprehensive performance of different methods under multi-objective optimization.

At the overall performance level, **Figure 1** shows the convergence behavior of different methods during training. It can be observed that the fixed-time method exhibits almost no learning process, and its performance curve remains largely stable. The DQN and PPO methods decrease rapidly in the first 100 episodes but then enter a slow optimization stage, showing obvious convergence lag. In contrast, the proposed method reaches a stable range after about 120 episodes, and its performance metrics decline significantly faster. From a numerical perspective, the performance value of the proposed method at the final convergence stage is

approximately 40, while PPO is approximately 45 and DQN approximately 50, representing improvements of about 11.1% and 20%, respectively. This result shows that with the introduction of the carbon emission perception mechanism and constraint optimization, policy learning efficiency is significantly improved, and the optimal solution can be approached more quickly.

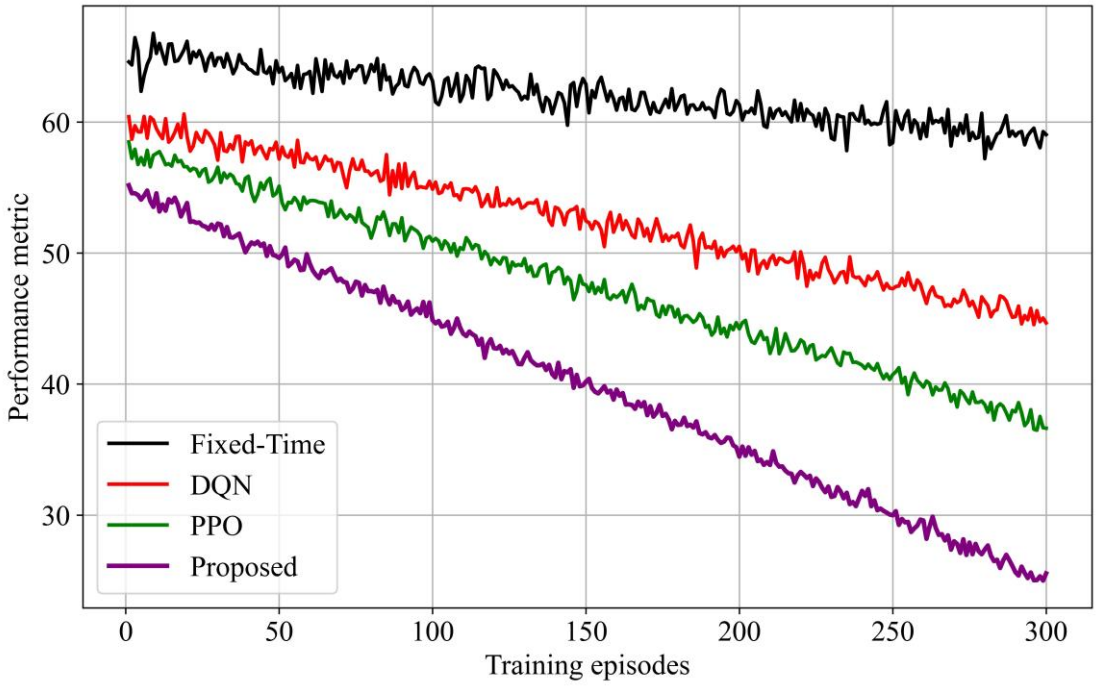

Figure 1. Training convergence curves of different methods

Further analysis of the slope of the convergence curve shows that the method in this paper exhibits the largest decline in the initial stage of training (0 – 100 rounds), indicating that it can quickly identify the high emission state and adjust the strategy early; In the later stage, the curve tends to be smooth, indicating that the strategy has converged stably. This "fast descent + steady convergence" characteristic is an important property of a high-quality reinforcement learning algorithm.

From a quantitative perspective, Table 2 shows the performance comparison of different methods across multiple metrics.

Table 2. Overall performance comparison of different methods

Method	Average discharge (g/s)	Incur loss through delay (s)	Average velocity (m/s)	Comprehensive indicators *J*
Fixed-time	52.3	38.7	6.2	63.8
DQN	44.2	28.4	7.8	49.3
PPO	41.6	26.7	8.3	45.2
Proposed	36.9	22.5	9.1	39.7
Unconstrained version	39.8	24.3	8.7	42.8

It can be observed from Table 2 that the proposed method achieves the best results across all core metrics. Among them, average carbon emissions decreased from 52.3 g/s for fixed-time

control to 36.9 g/s, a reduction of approximately 29.4%; average delay was reduced from 38.7 s to 22.5 s, an improvement of 41.9%; and average speed increased by approximately 46.8%. In terms of the comprehensive index J, the method in this paper achieves a further reduction by about 12.2% compared with PPO. In addition, the performance of the unconstrained version is approximately 7.8% lower than that of the full model, indicating that the constraint mechanism plays a key role in performance optimization. These results demonstrate that the proposed method achieves a better balance among multiple objectives.

For carbon emission reduction performance, we further analyze the emission intensity per vehicle:

$$I_E = \frac{E}{N} \tag{36}$$

Where E is the total emission and N is the total number of vehicles.

For carbon emission performance, Figure 2 shows the trend of unit emission intensity under different traffic intensities.

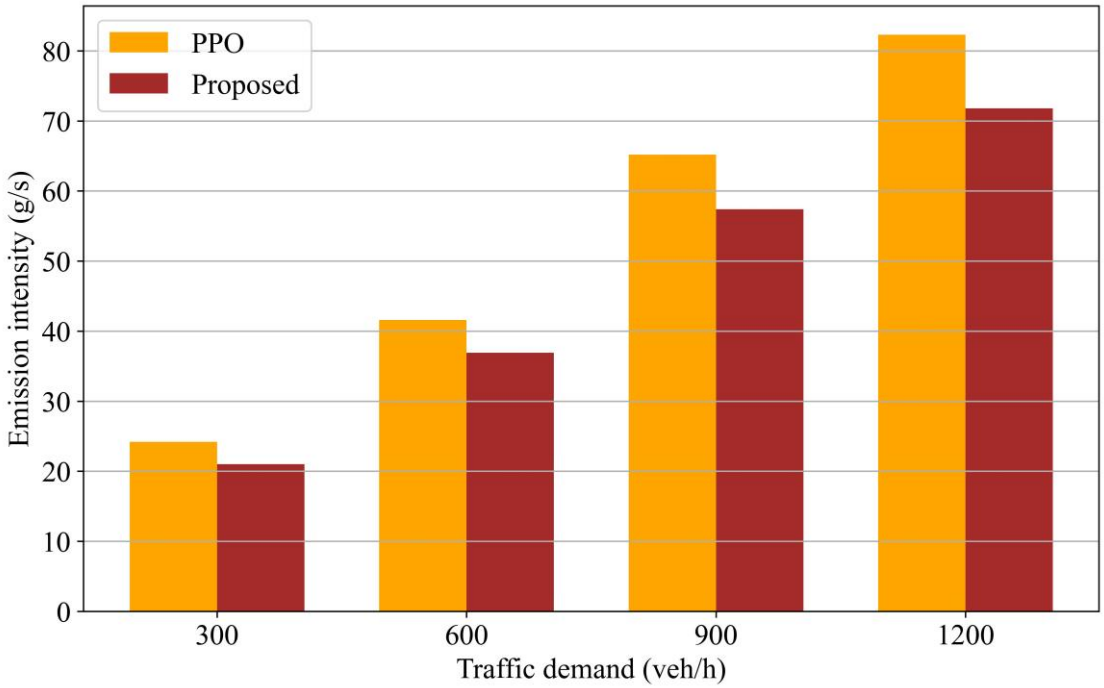

Figure 2. Comparison of unit emission intensity under different traffic intensity

It can be clearly seen that as traffic demand increases from 300 veh/h to 1200 veh/h, the emissions of each method show an upward trend, but the proposed method consistently maintains the lowest level. Under moderate load (600 veh/h), the emission of the proposed method is 36.9 g/s, which is approximately 11.3% lower than that of PPO (41.6 g/s). Under high load (1200 veh/h), the emission is 71.8 g/s, approximately 12.8% lower than PPO's 82.3 g/s. This result shows that the proposed method maintains stable emission control capability even under high-congestion conditions. For convergence and training stability, the policy volatility metric is defined as:

$$\sigma_\pi = \sqrt{\mathbb{E}[(J_t - \bar{J})^2]} \tag{37}$$

Where J_t is the performance value of episode t and $\bar{J}$ is the mean value. For training stability, Figure 3 shows the performance fluctuations of different methods during training. It can be observed that the PPO method exhibits obvious fluctuations, with a range of

approximately ±3, while the fluctuation range of the proposed method is controlled within ± 1.5.

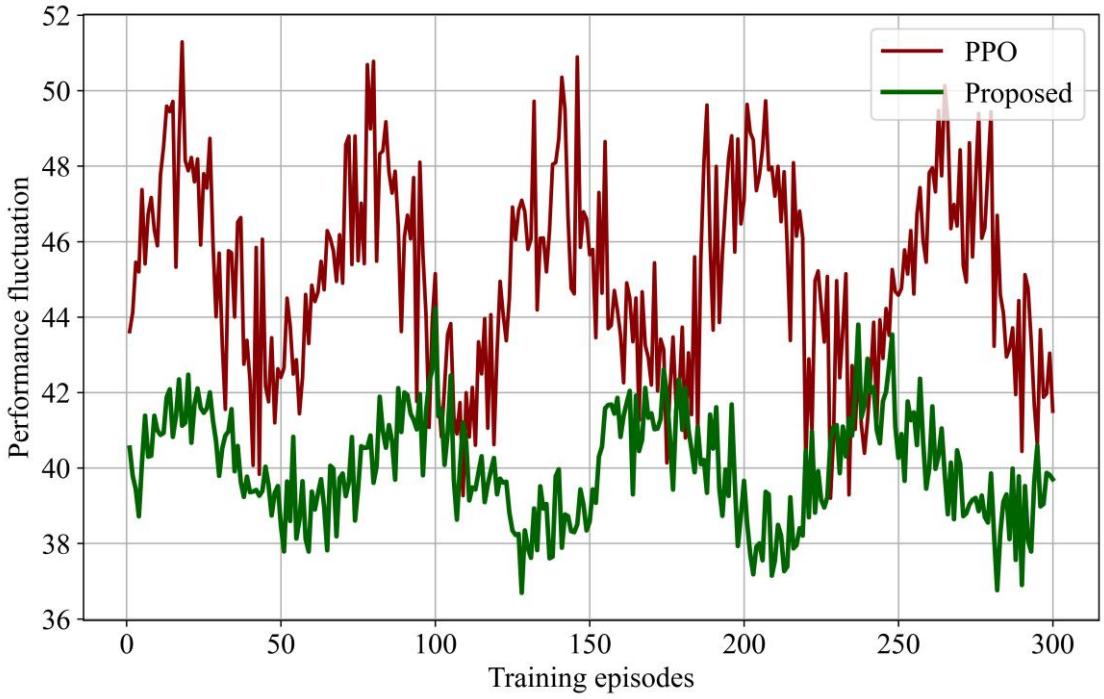

Figure 3. Comparison of stability of strategy training

The calculated value of PPO is $\sigma \approx 2.8$, while that of the proposed method is only $\sigma \approx 1.4$, representing a reduction in fluctuation by about 50%. This shows that the proposed dual value network and constraint mechanism significantly improve the training stability.

In the ablation study, the contribution of each key module was analyzed by gradually removing them. Table 3 shows the performance for different module combinations.

Table 3. Ablation study of different model components

Model version	Discharge (g/s)	Incur loss through delay (s)	Comprehensive indicators
Complete model	36.9	22.5	39.7
Unconstrained mechanism	39.8	24.3	42.8
No multiscale state	41.2	25.6	44.1
No dynamic rewards	43.5	27.2	46.3
Basic PPO	41.6	26.7	45.2

It can be seen from the results in Table 3 that the full model achieves the best performance in terms of emission and delay metrics. When the dynamic reward mechanism was removed, emissions increased from 36.9 g/s to 43.5 g/s, an increase of approximately 17.9%, indicating that this module has the greatest impact on carbon emission optimization. After removing multi-scale state modeling, delay increased by approximately 13.8%, indicating that spatial information is crucial for traffic efficiency optimization. After removing the constraint mechanism, emissions increased by approximately 7.9%. Overall, each module contributes significantly to performance improvement, with the dynamic reward mechanism and constraint

optimization being the core factors.

In the generalization ability test, the model is transferred to an unseen urban road network, and the generalization performance degradation rate is defined as:

$$R_g = \frac{J_{new} - J_{train}}{J_{train}} \times 100\% \tag{38}$$

Where J_{train} is the training scenario performance, and J_{new} is the new scene performance. The results are shown in Table 4.

Table 4. Generalization Performance on Unseen Traffic Scenarios

Method	Original scene J	New scene J	Generalization decline rate (%)
PPO	45.2	51.8	14.6
Proposed	39.7	43.5	9.6

As can be seen from Table 4, the comprehensive metric of the proposed method in the new scenario increased from 39.7 to 43.5, with a performance degradation rate of 9.6%, while the PPO method degraded by 14.6%. In contrast, the performance degradation of the proposed method is reduced by approximately 34.2%. This result shows that the proposed method has stronger generalization ability under different road network structures and traffic distribution conditions. In addition, the low rate of performance degradation indicates that the model learned more universal control strategies during training, rather than relying on specific scenario characteristics.

Finally, we analyze the policy behavior from the perspective of interpretability. We define the signal switching frequency as:

$$f_s = \frac{N_{switch}}{T} \tag{39}$$

Where N_{switch} is the number of signal switches. At the policy behavior level, Figure 4 shows the distribution of signal control strategies of different methods. It can be observed that the signal switching frequency of the proposed method is the lowest, at approximately 0.28, while those of DQN and PPO are 0.42 and 0.38, respectively.

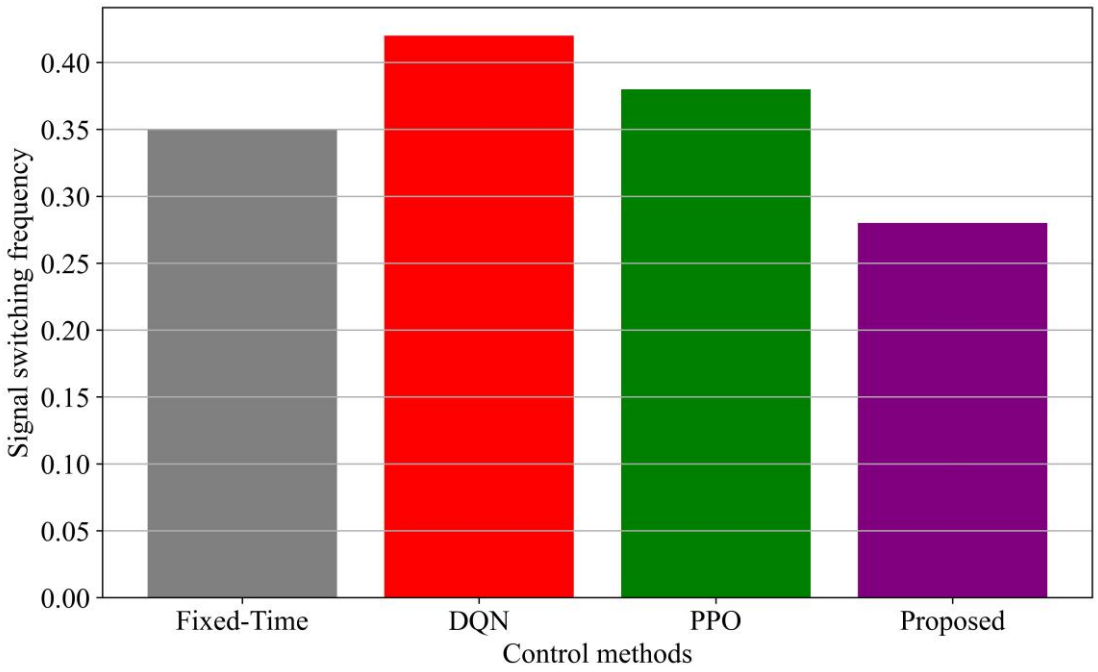

Figure 4. Distribution of signal control strategies

Combined with the emission results, it can be inferred that a lower signal switching frequency helps reduce vehicle acceleration and deceleration events, thereby reducing carbon emissions. This phenomenon shows that the model not only optimizes emission metrics numerically but also forms a decision-making model that aligns with physical principles at the policy level, exhibiting good interpretability.

In summary, through multi-dimensional experimental analysis, it can be confirmed that the proposed method outperforms existing methods in terms of carbon emission optimization, training stability, and generalization ability, with each key module contributing substantially to performance improvement.

7. DISCUSSION

The proposed deep reinforcement learning signal control algorithm for traffic carbon emission optimization demonstrates significant advantages in several aspects. From a methodological perspective, the proposed method systematically embeds carbon emission constraints into the reinforcement learning training process. Through the combination of a dynamic reward mechanism and the Lagrange multiplier method, it achieves an adaptive trade-off between emission objectives and traffic efficiency, overcoming the insufficient adaptability of traditional fixed-weight methods under varying traffic conditions. In terms of multi-intersection coordination, the distributed control framework based on parameter sharing and neighborhood information interaction effectively captures the spatial propagation characteristics of traffic flow, enabling local control decisions to account for the traffic status of upstream and downstream intersections, thereby improving the coordination capability of the overall road network. In terms of performance, the experimental results show that the proposed method outperforms existing baseline methods in the core metrics of carbon emissions, delay, and average speed, particularly in training stability and convergence speed. Its policy volatility is reduced by approximately 50% compared with traditional methods, and its generalization ability has also been verified in unseen scenarios. In addition, the model's learned behavior of low signal switching frequency aligns with the physical mechanism of reducing emissions from vehicle acceleration and deceleration, indicating that the algorithm not only achieves numerical optimization but also forms an interpretable control policy.

Although the proposed method achieves promising results, it still has some limitations. First, in terms of carbon emission estimation, this paper uses an instantaneous emission model based on speed and acceleration. Although it can reasonably describe the impact of vehicle operating status on emissions, it does not account for the differences in vehicle type, fuel type, and engine thermal state on emissions. In real traffic scenarios, there are significant differences in the emission characteristics of different vehicle types. If vehicle composition information can be incorporated into the modeling, it is expected to further improve the accuracy of emission estimation. Second, in terms of state representation, this paper mainly relies on traditional traffic parameters such as queue length, speed, and density, and has not fully utilized the lane-level and individual-level refined data available in the Internet of Vehicles environment, such as vehicle trajectories and signal phase timing, which limits the model's ability to perceive complex traffic situations to a certain extent. Third, from an algorithmic perspective, this paper uses a distributed control architecture in which each intersection agent makes decisions based on local states. Although information interaction is realized through the communication mechanism, the coordination complexity among multiple agents may increase significantly under high traffic congestion or further expansion of the road network scale, and policy convergence will face greater challenges. In addition, model training depends on a high-quality traffic simulation environment, and the gap between simulation and the real world may lead to performance degradation during migration and deployment.

Regarding the feasibility of actual deployment, several key factors for migration to real-world systems have been considered in the design and implementation of the proposed algorithm. From the perspective of computing resource requirements, the policy network adopts a lightweight fully connected structure, and the computational overhead of a single forward pass is small, which can meet the requirements of real-time signal control. In the training phase, although large-scale computing resources are required for simulation interaction, the training process can be completed in an offline environment. When the trained model is deployed at an actual intersection, only forward inference needs to be performed, and the computational demands on edge computing equipment are modest. From the perspective of interface compatibility, the control action output by the algorithm consists of phase selection and duration, which is compatible with the basic control commands of existing traffic signal controllers, eliminating the need to modify the underlying hardware. However, actual deployment still faces uncertainties in real-world traffic systems, sensor noise, and communication delays. In real-world environments, the integrity and real-time performance of vehicle detection data are difficult to achieve at the ideal level of simulation environments, which may lead to state estimation bias and subsequently affect policy effectiveness. In addition, signal control in real-world road networks often involves multi-department coordination, and changes to the control strategy must undergo a strict verification and approval process, which imposes higher requirements on the interpretability and reliability of the algorithm. Therefore, before actual deployment, it is necessary to combine edge computing, data fusion, security verification, and other technologies to build a full-chain trusted mechanism from training to deployment.

From the perspective of integration with ITS, the proposed method has the potential for collaborative development with a variety of emerging technologies. With the development of vehicle–road collaboration and autonomous driving technologies, information interaction between vehicles and infrastructure will become more real-time and accurate. In this context, the cooperative control mechanism based on neighborhood state encoding adopted in this paper can be further extended to a multi-source information fusion framework that integrates vehicle-level trajectory data. By obtaining real-time vehicle speed, position, and driving intention, signal timing can be optimized in advance, thereby further improving emission reduction performance. At the same time, with the in-depth application of digital twin technology in the transportation field, this algorithm can perform high-fidelity training and testing on a digital twin platform, reducing trial-and-error costs in real-world road networks through continuous iterative optimization based on virtual–real interaction. At the traffic control center level, the proposed method can serve as one of the core decision-making modules of the urban traffic

brain, operating in coordination with traffic prediction, route guidance, bus priority, and other functions to form a multi-level low-carbon traffic control system. In addition, the carbon emission constraint optimization framework proposed in this paper has a certain degree of universality and can be extended to other traffic management scenarios, such as ramp control, dynamic lane management, and regional traffic guidance, providing technical support for the development of low-carbon-oriented urban traffic systems. In summary, the proposed method not only achieves collaborative optimization of emissions and efficiency at the algorithmic level but also provides a feasible technical path for the evolution of intelligent transportation systems toward a greener and more efficient direction in the future.

8. CONCLUSION

Focusing on urban traffic carbon emission optimization, this paper systematically investigates signal control methods based on deep reinforcement learning. To address the shortcomings of existing research in carbon emission modeling, multi-objective conflict handling, and training stability, a deep reinforcement learning signal control algorithm for carbon emission optimization is proposed. Based on the coupling relationship between traffic congestion and carbon emissions, an intelligent signal control system integrating a microscopic emission estimation model, a dynamic reward mechanism, and a constraint optimization framework is constructed. The effectiveness and superiority of the proposed method are verified through simulation experiments from multiple perspectives. The results show that systematically embedding the carbon emission objective into the reinforcement learning training process can significantly reduce carbon emissions while accounting for traffic efficiency, providing a new solution for green traffic control.

The main contributions of this paper are fourfold. First, at the problem modeling level, a multi-objective carbon-emission-aware reward mechanism is proposed. Through dynamic weight adjustment, the model can adaptively balance the trade-off between emissions and efficiency according to traffic conditions, effectively alleviating the insufficient adaptability of traditional fixed-weight methods. Second, at the algorithm design level, constrained reinforcement learning is introduced into the field of traffic signal control. The Lagrange multiplier method is used to explicitly integrate the carbon emission threshold into the policy optimization process, and the dual value network, prioritized experience replay, and adaptive exploration strategy are combined to significantly improve training stability and convergence efficiency. Third, at the system architecture level, a distributed collaborative control mechanism based on parameter sharing and neighborhood information interaction is designed for multi-intersection scenarios, which enhances the model's ability to perceive the spatial propagation characteristics of traffic flow and achieves unification of local decision-making and global coordination. Fourth, at the experimental validation level, a multi-scenario test environment covering a single intersection, a grid network, and a real-world urban road network was constructed, and the method was comprehensively evaluated across multiple dimensions, such as overall performance, convergence characteristics, generalization ability, and policy interpretability, which verified its significant advantages in reducing carbon emissions, improving delay, and enhancing training stability.

Future research can expand this study in the following directions. First, in terms of emission modeling, the current microscopic emission model is mainly based on speed and acceleration parameters, which fails to fully reflect the impact of vehicle type, fuel composition, road grade, and other factors on emissions. In future research, a more refined emission model can be introduced. Combined with vehicle identification data and onboard diagnostic systems, a personalized, multi-granularity carbon emission estimation method can be constructed to further improve optimization accuracy. Second, in terms of state perception, with the development of the Internet of Vehicles and autonomous driving technologies, the dimension and quality of data available in traffic environments will continue to improve. In the future, we can explore the integration of multi-source information such as vehicle trajectories, signal status,

and roadside perception, and use advanced architectures such as graph neural networks and attention mechanisms to enhance the agent's ability to represent complex traffic situations and achieve more accurate control decisions. Third, in terms of multi-agent collaboration, the current distributed framework may face the problem of declining coordination efficiency as the scale of the road network expands further. In the future, we can study hierarchical control architectures, combine regional coordination with local intersection optimization, and explore agent collaboration mechanisms based on communication efficiency optimization to improve the scalability of large-scale road networks. Fourth, in terms of actual deployment, the gap between simulation environments and the real world remains an important factor limiting the implementation of the algorithm. In the future, we can combine digital twin technology to build a high-fidelity simulation platform, conduct transfer learning research on virtual – real interaction, and introduce security verification and robustness enhancement mechanisms to improve the reliability and trustworthiness of the algorithm in real-world traffic environments. In addition, expanding this method to other low-carbon traffic management scenarios, such as bus priority signal control, eco-driving guidance, and dynamic lane management, to build an integrated urban traffic carbon emission control system is also a valuable research direction.

Abbreviations

MDP, Markov Decision Process;
SUMO, Simulation of Urban Mobility;
TraCI, Traffic Control Interface;
DQN, Deep Q-Network;
PPO, Proximal Policy Optimization;
MAE, Mean Absolute Error;
RMSE, Root Mean Square Error;
TD, Temporal Difference;
GPU, Graphics Processing Unit;
CPU, Central Processing Unit;
ITS, Intelligent Transportation System;
IMU, Inertial Measurement Unit;
CAV, Connected and Autonomous Vehicle;
VSP, Vehicle Specific Power;
OBD, On-Board Diagnostics;
GNN, Graph Neural Network;
CNN, Convolutional Neural Network;
LSTM, Long Short-Term Memory.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **H.X.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **H.X.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Li, B. W., Chen, Z. H., Zhu, X. H., Zhang, Z., Peng, Z. R., Zhao, H. M., & He, H. D. (2025). Assessment of eco-driving strategies on carbon emissions for hybrid vehicles through portable emissions measurement systems. *Atmospheric Pollution Research*, *16*(3), 102365. DOI: **https://doi.org/10.1016/j.apr.2024.102365**

[2] Chavhan, S., Deepika, I. S., Gupta, D., & Rodrigues, J. J. (2025). Energy-Efficient-Enabled Edge-AI-IoT Integrated Traffic Incident Analysis and Avoidance of Secondary Incidents. *IEEE Internet of Things Journal*. DOI: **https://doi.org/10.1109/JIOT.2025.3555408**

[3] Li, X., Wang, G., Zhu, Y., & Liu, W. (2025). A System Dynamics-Based Simulation Study on Urban Traffic Congestion Mitigation and Emission Reduction Policies. *Sustainability*, *17*(20), 9296. DOI: **https://doi.org/10.3390/su17209296**

[4] Li, D., Zhu, F., Wu, J., Wong, Y. D., & Chen, T. (2024). Managing mixed traffic at signalized intersections: An adaptive signal control and CAV coordination system based on deep reinforcement learning. *Expert Systems with Applications*, *238*, 121959. DOI: **https://doi.org/10.1016/j.eswa.2023.121959**

[5] Benhamza, K., Seridi, H., Agguini, M., & Bentagine, A. (2024). A multi-agent reinforcement learning based approach for intelligent traffic signal control. *Evolving Systems*, *15*(6), 2383-2397. DOI: **https://doi.org/10.1007/s12530-024-09622-4**

[6] Chen, X., Wang, X., Zhao, W., Wang, C., Cheng, S., & Luan, Z. (2025). Hierarchical deep reinforcement learning based multi-agent game control for energy consumption and traffic efficiency improving of autonomous vehicles. *Energy*, *323*, 135669. DOI: **https://doi.org/10.1016/j.energy.2025.135669**

[7] Hu, J., Shan, Y., Yang, Y., Parisio, A., Li, Y., Amjady, N., ... & Rodríguez, J. (2023). Economic model predictive control for microgrid optimization: A review. *IEEE Transactions on Smart Grid*, *15*(1), 472-484. DOI: **https://doi.org/10.1109/TSG.2023.3266253**

[8] Qadri, S. S. S. M., Gökçe, M. A., & Öner, E. (2020). State-of-art review of traffic signal control methods: challenges and opportunities. *European transport research review*, *12*(1), 55. DOI: **https://doi.org/10.1186/s12544-020-00439-1**

[9] Tedjopurnomo, D. A., Bao, Z., Zheng, B., Choudhury, F. M., & Qin, A. K. (2020). A survey on modern deep neural network for traffic prediction: Trends, methods and challenges. *IEEE Transactions on Knowledge and Data Engineering*, *34*(4), 1544-1561. DOI: **https://doi.org/10.1109/TKDE.2020.3001195**

[10] Liu, Y., Lyu, C., Zhang, Y., Liu, Z., Yu, W., & Qu, X. (2021). DeepTSP: Deep traffic state prediction model based on large-scale empirical data. *Communications in transportation research*, *1*, 100012. DOI: **https://doi.org/10.1016/j.commtr.2021.100012**

[11] Luo, R., Peng, Z., & Hu, J. (2023). On model identification based optimal control and it's applications to multi-agent learning and control. *Mathematics*, *11*(4), 906. DOI: **https://doi.org/10.3390/math11040906**

[12] Nguyen, T. T., Nguyen, N. D., & Nahavandi, S. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, *50*(9), 3826-3839. DOI: **https://doi.org/10.1109/TCYB.2020.2977374**

[13] Liu, H., Li, X., Zhang, L., & Cheng, R. (2026). Bridging phase and timing: A joint Q-value learning framework for synergistic traffic signal control at consecutive arterial road intersections. *Physica A: Statistical Mechanics and its Applications*, 131421. DOI: **https://doi.org/10.1016/j.physa.2026**

[14] Bernárdez, G., Suárez-Varela, J., López, A., Shi, X., Xiao, S., Cheng, X., ... & Cabellos-Aparicio, A. (2023). Magnneto: A graph neural network-based multi-agent system for traffic engineering. *IEEE Transactions on Cognitive Communications and Networking*, *9*(2), 494-506. DOI: **https://doi.org/10.1109/TCCN.2023.3235719**

[15] Wang, X., Yue, X., Huang, J., & Li, S. (2025). Integrating traffic dynamics and emissions modeling: From classical approaches to data-driven futures. *Atmosphere*, *16*(6), 695. DOI: **https://doi.org/10.3390/atmos16060695**

[16] Mera, Z., Varella, R., Baptista, P., Duarte, G., & Rosero, F. (2022). Including engine data for energy and pollutants assessment into the vehicle specific power methodology. *Applied Energy*, *311*, 118690. DOI: **https://doi.org/10.1016/j.apenergy.2022.118690**

[17] He, K., Chen, C., Chen, S., Chen, B., Zhang, A., Chen, P., ... & Wu, Z. (2025). Reinforcement Learning for Multi-Objective Optimization: A Review. *Archives of Computational Methods in Engineering*, 1-30. DOI: **https://doi.org/10.1007/s11831-025-10389-3**

[18] Nguyen, T. T., Nguyen, N. D., Vamplew, P., Nahavandi, S., Dazeley, R., & Lim, C. P. (2020). A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence*, *96*, 103915. DOI: **https://doi.org/10.1016/j.engappai.2020.103915**

[19] Liu, X., Ye, K., van Vlijmen, H. W., Emmerich, M. T., IJzerman, A. P., & van Westen, G. J. (2021). DrugEx v2: de novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. *Journal of cheminformatics*, *13*(1), 85. DOI: **https://doi.org/10.1186/s13321-021-00561-9**

[20] Pereira, V., Sousa, P., & Rocha, M. (2022). A comparison of multi-objective optimization algorithms for weight setting problems in traffic engineering. *Natural Computing*, *21*(3), 507-522. DOI: **https://doi.org/10.1007/s11047-020-09807-1**

[21] Taha, K. (2020). Methods that optimize multi-objective problems: A survey and experimental evaluation. *IEEE Access*, *8*, 80855-80878. DOI: **https://doi.org/10.1109/ACCESS.2020.2989219**

[22] Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 11216-11235. DOI: **https://doi.org/10.1109/TPAMI.2024.3457538**

[23] Ceusters, G., Camargo, L. R., Franke, R., Nowé, A., & Messagie, M. (2023). Safe reinforcement learning for multi-energy management systems with known constraint functions. *Energy and AI*, *12*, 100227. DOI: **https://doi.org/10.1016/j.egyai.2022.100227**

[24] Motte, M., & Pham, H. (2022). Mean-field Markov decision processes with common noise and open-loop controls. *The Annals of Applied Probability*, *32*(2), 1421-1458. DOI: **https://doi.org/10.1214/21-AAP1713**

[25] Yang, J., Wu, J., Fang, L., Fan, H., Zhang, B., Zhao, H., ... & You, X. (2025). MSRFormer: road network representation learning using multi-scale feature fusion of heterogeneous spatial interactions. *Geo-spatial Information Science*, 1-20. DOI: **https://doi.org/10.1080/10095020.2025.2583710**

[26] Ye, C., Liu, F., Ou, Y., & Xu, Z. (2022). Optimization of Vehicle Paths considering Carbon Emissions in a Time-Varying Road Network. *Journal of advanced transportation*, *2022*(1), 9656262. DOI: **https://doi.org/10.1155/2022/9656262**

[27] Li, H., Qian, X., & Song, W. (2024). Prioritized experience replay based on dynamics priority. *Scientific Reports*, *14*(1), 6014. DOI: **https://doi.org/10.1038/s41598-024-56673-3**

[28] Vadlamani, S. K., Xiao, T. P., & Yablonovitch, E. (2020). Physics successfully implements Lagrange multiplier optimization. *Proceedings of the National Academy of Sciences*, *117*(43), 26639-26650. DOI: **https://doi.org/10.1073/pnas.2015192117**

[29] Saeed Chilmeran, H. T., Hamed, E. T., Ahmed, H. I., & Al-Bayati, A. Y. (2022). A method of two new augmented lagrange multiplier versions for solving constrained problems. *International journal of mathematics and mathematical sciences*, *2022*(1), 3527623. DOI: **https://doi.org/10.1155/2022/3527623**

[30] Chen, R., Tsay, Y. S., & Ni, S. (2022). An integrated framework for multi-objective optimization of building performance: Carbon emissions, thermal comfort, and global cost. *Journal of Cleaner Production*, *359*, 131978. DOI: **https://doi.org/10.1016/j.jclepro.2022.131978**

[31] Zubaer, K. H., Alam, Q. M., Toha, T. R., Salim, S. I., & Al Islam, A. A. (2020). Towards simulating non-lane based heterogeneous road traffic of less developed countries using authoritative polygonal GIS map. *Simulation Modelling Practice and Theory*, *105*, 102156. DOI: **https://doi.org/10.1016/j.simpat.2020.102156**

[32] Chen, D., Zhu, M., Yang, H., Wang, X., & Wang, Y. (2024). Data-driven traffic simulation: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, *9*(4), 4730-4748. DOI: **https://doi.org/10.1109/TIV.2024.3367919**